DOCUMENT RESUME

ED 422 393 TM 028 955

AUTHOR Lee, Guemin; Frisbie, David A.
TITLE A Generalizability Approach To Evaluating the Reliability of Testlet-Based Test Scores.
PUB DATE 1997-03-26
NOTE 35p.; Paper presented at the Annual Meeting of the American Educational Research Association (Chicago, IL, March 24-28, 1997).
PUB TYPE Numerical/Quantitative Data (110) -- Reports - Evaluative (142) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Estimation (Mathematics); *Generalizability Theory; *Reliability; *Scores; Tables (Data); *Test Items; Test Results
IDENTIFIERS *Testlets

ABSTRACT

Previous studies have indicated that the reliability of test scores composed of testlets might be overestimated by conventional item-based reliability estimation methods (R. Thorndike, 1953; A. Anastasi, 1988; S. Sireci, D. Thissen, and H. Wainer, 1991; H. Wainer and D. Thissen, 1996). This study used generalizability theory to investigate the relative adequacy of reliability coefficients from test scores composed of testlets with a p x (I:H) random effects design, where persons are crossed with items nested within passages. The magnitude of overestimation of using Cronbach's coefficient alpha based on item scores in this situation was estimated to be about 0.04. The passage facet turns out to be more influential on reliability estimates than the item-within-passage facet. Given a fixed total number of items and a fixed number of passages, the variability of generalizability coefficients with varying number of items per passage is small (under 0.01). Therefore, manipulating the number of passages is a more productive way to obtain efficient measurement procedures than is manipulating the number of items within each passage. (Contains 7 tables, 3 figures, and 18 references.) (Author)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 422 393

A Generalizability Approach to Evaluating the Reliability of Testlet-based Test Scores

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Guemin Lee

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

1

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

Guemin Lee
David A. Frisbie
University of Iowa

Paper presented at the 1997 Annual Meeting of the National Council on Measurement in Education Chicago, IL March 26, 1997

TM028955

A Generalizability Approach to Evaluating the Reliability of Testlet-based Test Scores

Abstract

Previous studies have indicated that the reliability of test scores composed of testlets might be overestimated by conventional item-based reliability estimation methods (Thorndike, 1951; Anastasi, 1988; Sireci, Thissen & Wainer, 1991; Wainer & Thissen, 1996). This study used generalizability theory to investigate the relative adequacy of reliability coefficients from test scores composed of testlets with a p x (I : H) random effects design, where persons are crossed with items nested within passages. The magnitude of overestimation of using Cronbach's coefficient ALPHA based on item scores in this situation was estimated to be about 0.04. The passage facet turns out to be more influential on reliability estimates than the item-within-passage facet. Given a fixed total number of items and a fixed number of passages, the variability of generalizability coefficients with varying number of items per passage is small (under 0.01). Therefore, manipulating the number of passages is a more productive way to obtain efficient measurement procedures than is manipulating the number of items within each passage.

A Generalizability Approach to Evaluating the Reliability of Testlet-based Test Scores

Introduction

"Testlets" are small tests, small enough to manipulate but large enough to carry their own context (Wainer & Lewis, 1990; Wainer & Kiely, 1987). The focus of this paper is on the previous research finding that the reliability of test scores obtained from testlets generally will be overestimated when item-based reliability estimation methods are used (Thorndike, 1951; Anastasi, 1988; Sireci, Thissen & Wainer, 1991; Wainer & Thissen, 1996). One common definition of reliability is based on the idea that the same test results should be obtained with equivalent measures. On the basis of this definition, if internal consistency reliability coefficients are computed properly, they will accurately estimate the corresponding equivalent forms correlations. However, when some items in a test are related to the same single passage or other stimulus material, there is dependence among those items, and internal consistency estimates of reliability might be inflated relative to estimates of reliability based on the correlation between equivalent forms of the test (Lawrence, 1995). The purpose of this study is to investigate the adequacy of various reliability estimates of testlet-based test scores.

According to the *Test Standards* (AERA, APA, NCME, 1985), obtaining and reporting evidence concerning reliability and errors of measurement are the fundamental responsibilities of test developers and publishers. Such evidence on the uncertainty attached to group and individual scores is required to avoid overinterpretation of scores (Cronbach, Linn, Brennan & Haertel, 1995). If the reliability of test scores obtained from testlets were overestimated by the item-based reliability estimation methods, these estimates might lead to the misinterpretation of scores--treating scores as though they are more consistent than they actually are.

Because there is little evidence in the literature about how large the reliability overestimates might be in this situation, it is not clear how serious the score misinterpretation might be. This study was designed to permit comparisons among estimates that would inform users about how serious the overestimation problem might be for practical score interpretation purposes.

This problem can be addressed by considering four possible methodological approaches: Cronbach's coefficient ALPHA, stratified coefficient ALPHA, item response theory (IRT), and generalizability theory (G-theory). If the passages are treated as a fixed factor, stratified coefficient ALPHA can be used to estimate reliability. But, if the passages are considered a random factor, as is nearly always the case, stratified coefficient ALPHA is inappropriate. Therefore, this study has not included stratified coefficient ALPHA.

The use of Cronbach's coefficient ALPHA depends on the assumption that the part scores (or item scores) are essentially tau-equivalent (Feldt & Brennan, 1989). If the average inter-item correlation within testlets exceeds the average inter-item correlations between testlets, this assumption would be violated. That is, the presence of a systematic pattern of inter-item correlations could violate the assumption. If the level of dependence within passages is found to be relatively higher than that between passages, the passage scores would be the most appropriate unit of analysis for estimating reliability (Frisbie & Druva, 1986). In this paper, two types of coefficient ALPHA are distinguished: Item α is based on item scores and Passage α is based on testlet or passage scores. (Passage scores can be calculated by summing up the item scores within each passage.)

Wainer & Thissen (1996) and Sireci, Thissen, & Wainer (1991) studied this topic using IRT approaches and concluded that the overestimation is due to "local dependence". The presence of conditional dependence, a seemingly natural by-product when some items have a common stimulus, implies that the items from the

total test measure more than one construct. When using the approaches, the researcher should be cautious about two important points. First, the researcher should consider the consequences of the particular scoring method selected. According to Wainer and Lewis (1990), the items of a test composed of testlets usually violate the assumption of conditional independence among items. These authors suggested three alternative responses to treat this problem : 1) modify the number of items so that each passage has only a single item, 2) ignore the interdependencies among the items and fit a binary response model, 3) define the passage with its associated questions as a single item. For this third approach, Sireci, Thissen, & Wainer (1991) and Wainer & Thissen (1996) used Bock's model in which the researcher treats the examinee's responses to the m passages as the responses to m polychotomous items, and then scores them either 0, 1, 2, ..., or m. If the researcher were to use a different scoring scheme from Bock's model, he/she would get different results. Second, the IRT approach requires strong assumptions. That is, in order to apply the IRT approach to this situation, the researcher must provide evidence that the IRT assumptions (e.g., dimensionality and local independence) have been satisfied.

A G-theory approach could avoid the above problems of using IRT approaches. That is, with G-theory, there would be no concern about the different scoring methods. Furthermore, G-theory is considered a "weak theory", which means it doesn't require any strong assumptions. In addition to these two important advantages, G-theory requires less computer time and effort, and it may be conceptually more understandable and straight-forward for practitioners.

The univariate p x (I : H) D-study design, persons crossed with items nested in passages, is appropriate for this study. Assuming a balanced design, which means the number of items within passages is equal, the generalizability coefficient can be computed by Equation 1. The term $\sigma^2(p)$ represents the universe score variance, and

$\sigma^2(\delta)$ can be defined as the relative error variance. The term $\sigma^2(pH)$ represents the person by passage interaction variance component in a D-Study. Similarly, the term $\sigma^2(pI{:}H)$ can be defined as the variance component in a D-Study, representing the persons by items within a given passage interaction.

$$E\rho^2 = \frac{\sigma^2(p)}{\sigma^2(p)+\sigma^2(\delta)} \quad \text{where} \quad \sigma^2(\delta) = \sigma^2(pH)+\sigma^2(pI{:}H) \quad \text{----------} \quad (1)$$

Traditional reliability estimation methods like coefficient ALPHA treat the passage facet as a "hidden fixed facet". In this case, the formula for computing the generalizability coefficient is defined by Equation 2. The term $\sigma^2(\tau)$ represents the universe score variance, which is composed of $\sigma^2(p)$ and $\sigma^2(pH)$ and $\sigma^2(\delta)$ is $\sigma^2(pI{:}H)$.

$$E\rho^2 = \frac{\sigma^2(\tau)}{\sigma^2(\tau)+\sigma^2(\delta)} \quad \text{where} \quad \sigma^2(\tau) = \sigma^2(p)+\sigma^2(pH) \text{ and } \sigma^2(\delta) = \sigma^2(pI{:}H) \quad \text{-- (2)}$$

From a comparison of Equations 1 and 2, it can be seen that the $\sigma^2(pH)$ term contributes to relative error variance in Equation 1, but it contributes to the universe score variance in Equation 2. Therefore, in a given situation, the generalizability coefficient from Equation 2 will be greater than that from Equation 1. The traditional reliability coefficients are analogous to Equation 2. Thus, it can be seen that the reliability of test scores built from testlets may be overestimated by using conventional item-based reliability estimation methods.

The conditions for a balanced design are not common in practice because usually the number of items per passage varies among passages. For an unbalanced design, there are a number of procedures reported in the literature for estimating variance components in a G-study (Brennan, 1994). Jarjoura & Brennan (1981) provided the ANOVA-like procedures for estimating variance components for the random effects p x (i : h) G-study design with unequal number of items per passage. If the numbers of items per passage are independent of the random effects in the model, then the estimators of the variance components are unbiased (Brennan,

Jarjoura, & Deaton, 1980; Jarjoura & Brennan, 1981; Brennan, 1992). ANOVA-like procedures for estimating variance components were used in this study.

This study has three primary objectives:

1. Investigate the size of the difference among Item α, Passage α, and the generalizability coefficient for each test and each grade level.

2. Examine the difference among reliability estimates by doing an analysis of the random variables created by the within-passage and between-passage inter-item correlations.

3. Determine the influence of the number of testlets and the number of items within each testlet on the generalizability coefficients and on the size of the difference between ALPHA and the generalizability coefficient.

Method

Data Sources

The data for this study were taken from the spring 1992 *Iowa Tests of Basic skills* (ITBS) and *Iowa Tests of Educational Development* (ITED) national standardization sample for Form K. In this study, grade 4, 8, and 11 students were used because the test structures used in these three grades are considered representative of grade levels 3-12. A 30% random sample was selected from the national standardization sample for grade 4 and grade 8, and the whole national standardization sample was taken for grade 11. The sample size and the general characteristics of each test are presented in Table 1.

Insert Table 1 About Here

The tests used are the Reading Comprehension and Maps and Diagrams tests of the ITBS for grades 4 and 8 and Test L: Ability to Interpret Literary Materials of the ITED for grade 11. The Reading Comprehension test measures how well students can

comprehend a variety of written materials. There are nine passages in each test level. The number of items per passage ranges from two to six for grade 4 and three to twelve for grade 8. .(The Reading Comprehension items from Form K used in this study are slightly different from the operational version of Form K.) The skills measured by the Maps and Diagrams test are process-oriented: students must apply their skills to visuals such as maps, diagrams, and charts, none of which they have ever seen before. There are four or five maps, diagrams, and charts, each with six to seven items, in each test level (Hoover, Hieronymus, Frisbie, & Dunbar, 1994). In Test L, there are five selections including about nine items per passage at each test level. The excerpts are from novels, short stories, memoirs, and essays, and they range in length from 275 to 700 words (Feldt, Forsyth, Ansley, & Alnot, 1994).

Design

A linear model for the response of a person to an item within a passage was used for this study. Persons are objects of measurement, and items and passages are treated as random facets. For this model, n_p persons represent a random sample from a population of interest and n_h passages represent a random sample from the universe of passages. The $n_{i:h}$ items in a passage are also considered as a random sample from that passage that are selected independently of other passages. This linear model, referred to as completely random, can be represented as in Equation 3.

$$
\begin{aligned}
X_{pih} &= \mu && \text{(grand mean)} \quad \text{---------- (3)} \\
&+ \mu_p - \mu && \text{(person effect)} \\
&+ \mu_h - \mu && \text{(passage effect)} \\
&+ \mu_{i:h} - \mu_h && \text{(item within passage effect)} \\
&+ \mu_{ph} - \mu_p - \mu_h + \mu && \text{(person by passage interaction effect)} \\
&+ X_{pi:h} - \mu_{ph} - \mu_{i:h} + \mu_h && \text{(residual effect)}
\end{aligned}
$$

where p=1, ... , n_p, i=1, ... , $n_{i:h}$, h=1, ... , n_h

In G-theory, the generalizability of a particular measurement procedure

depends upon how the scores will be used in making decisions. Two different types of error variances are associated with separate types of decisions: relative and absolute decisions (Cronbach, Gleser, Nanda, & Rajaratnam, 1972; Brennan, 1992). In this study, the major interest is in relative decisions in order to make comparisons with the conventional reliability coefficients based on individual items. In this paper, n_+ is the total number of items in a test (or in a D-study), which may or may not be the same as that of original test (or in a G-study). The relative error variance of this random effects p x (I : H) D-study design can be calculated by Equation 4, and the generalizability coefficient can be obtained using Equation 1, which represents the ratio of the universe score variance to the observed score variance that is composed of the universe score variance and relative error variance (Jajoura & Brennan, 1981).

$$\sigma^2(\delta) = \frac{1}{n_+}[V_i \times \sigma^2(ph) + \sigma^2(pi{:}h)] \quad \text{where } V_i = \frac{\sum_h n_{i:h}^2}{n_+} \qquad \text{---------- (4)}$$

Analyses

The traditional reliability coefficients, based on individual items and then on passage scores, were computed using Cronbach's ALPHA coefficient. The G-study analysis was conducted using ANOVA-like procedures in order to estimate variance components.[1] Then, in several D-studies, which have the purpose of determining the most efficient measurement procedure, the generalizability coefficients were computed for the same measurement structure as in the G-study. For the first research question, reliability estimates are compared among Item α, Passage α, and generalizability coefficients for each test and each grade level. These results furnish the empirical data about how much the reliability estimates of testlet-based test scores are overestimated by item-based reliability estimation methods.

To explain the differences among reliability estimates, five random variables for within-passage inter-item correlations and five random variables for between-

passage inter-item correlations were constructed. The distributional characteristics of the two random variables (one for within passage and one for between passage) for each test and for each grade level were then compared.

To examine the factors influencing the testlet-based reliability estimates, a variety of D-studies were done by manipulating the number of passages and the number of items within each passage. The D-studies were constructed for the same universe of generalization as the universe of admissible observation in the G-studies, and all D-studies were conducted under a complete random effects p x (I : H) design. For comparing the conventional reliability estimates with the various kinds of D-Study results, the generalizability coefficients for the p x I random effects design (exactly the same as the Cronbach's coefficient ALPHA based on item scores) were computed, and these were compared with the generalizability coefficients of the p x (I : H) random effects design.

Results and Discussion

Table 2 provides reliability estimates from Item α, Passage α, and the generalizability coefficient, indicating that Cronbach ALPHA coefficient based on item scores is higher than both the generalizability coefficient and the Cronbach ALPHA coefficient based on passage scores.

Insert Table 2 About Here

The average difference between Item α and the generalizability coefficient is about .040. This difference can be explained by the fact that Cronbach's ALPHA coefficient based on item scores ignores the passage facet. That is, as noted in the introduction section, the variance component, $\sigma^2(pH)$, contributes to true score variance (or universe score variance) in calculating Cronbach's ALPHA coefficient based on item scores, but it contributes to relative error variance in calculating the

generalizability coefficient of the random effects p x (I : H) D-study design.

The practical effect of the difference between reliability estimates can be shown by using a confidence interval around a raw score (using standard error of measurement). Suppose a certain grade 8 student gets a raw score 29 on the ITBS Maps and Diagrams test. That student's raw score confidence interval, with one S.E.M., is 25.70 to 32.30 (or rounded, 26 to 32) using Cronbach's coefficient ALPHA based on item scores. However, it is 24.89 to 33.11 (or rounded, 25 to 33) using the generalizability coefficient. The confidence interval around a raw score using the generalizability coefficient is a little bit wider than that using Cronbach's coefficient ALPHA based on item scores. However, the difference of confidence intervals based on the two different reliability estimates is very small, and it doesn't seem to lead to the serious misinterpretation of the scores in a practical sense.

In order to explain the difference between Item α and Passage α, the relationship of Cronbach's coefficient ALPHA to the Spearman-Brown formula can be analyzed. Cronbach's coefficient ALPHA can be obtained from the Spearman-Brown formula by replacing the correlation coefficient by the average of the item covariance divided by the average test variance ($_{\alpha}\rho_{xx'} = \frac{n\overline{\sigma}_{X_f X_g}}{\overline{\sigma}^2_{X_f} + (n-1)\overline{\sigma}_{X_f X_g}}$). If the multiple parts of the test are the classically parallel forms (or items), this formula is exactly the same as the Spearman-Brown formula. The purpose here is to provide an explanation for higher reliability estimates with Item α than with other reliability estimates. Item α can be approximated by applying the Spearman-Brown formula after obtaining the average of all inter-item correlations. First, denote $\overline{\rho}_w$ as the average of within-passage inter-item correlations and $\overline{\rho}_b$ as the average of between-passage inter-item correlations, and $\overline{\rho}_t$ as the average of all inter-item correlations. Item α can be approximated by using this formula $\rho = \frac{n\overline{\rho}_t}{1+(n-1)\overline{\rho}_t}$.

Consider $\overline{\rho}_t$ as the weighted average of the $\overline{\rho}_w$ and $\overline{\rho}_b$. If $\overline{\rho}_w$ is equal to $\overline{\rho}_b$, then the appropriate reliability estimate is obtained, but, if $\overline{\rho}_w$ is greater than $\overline{\rho}_b$, indicating a violation of the assumption of Cronbach's ALPHA coefficient, a biased reliability estimate would result. According to Table 2, the average difference between Item α and Passage α for these data is about .048. This difference can be explained by the difference between the average within-passage inter-item correlations and the average between-passage inter-item correlations, which are shown in Table 3.

Insert Table 3 About Here

The average within-passage inter-item correlations range from .170 to .266 and the average between-passage inter-item correlations range from .128 to .216 for the five tests in this study. The average of within-passage inter-item correlations is 1.34 times greater than that of between-passage inter-item correlations. Relative to a normal distribution, the within-passage and between-passage inter-item correlations have similar distributional forms, a little positively skewed, except for the ITBS grade 8 Maps and Diagrams data, and a little leptokurtic, except for the ITBS grade 4 Maps and Diagrams and ITED grade 11 Test L data. Therefore, the two distributions of within and between passage inter-item correlations are different from each other, especially in their location statistics. Under these circumstances, Frisbie & Durva (1986) recommended the use of passage scores instead of item scores to eliminate the dependence among within passage items. The wisdom of this advice can be judged in part by the data in Table 2. The difference between Item α and Passage α is about .048, and systematically greater than the difference between Item α and the generalizability coefficient. And the average difference between Passage α and the generalizability coefficient is about .008, a very small difference.

To examine the factors influencing the reliability estimates of the p x (I : H)

random effects design, a variety of D-studies were completed. The D-studies were of two types, one for a fixed total number of items and the other for a varied total number of items. The D-studies with a varied total number of items dealt with the relative importance of the passage effect and the item-within-passage effect. The D-studies with a fixed total number of items dealt with the confounded effect of the passages and items within each passage. Table 4 and Figure 1 provide the generalizability coefficients of the p x (I : H) random effects D-study design with varying number of passages and varying numbers of items per passage.

Insert Table 4 About Here

Insert Figure 1 About Here

The generalizability coefficients increase at a greater rate by increasing the number of passages than by increasing the number of items per passage. This generalization can be confirmed by the following example with data from Table 4.

H'=3, I'=4, Total n=12, G-coeff.=.60812

H'=4, I'=4, Total n=16, G-coeff.=.67417

H'=3, I'=5, Total n=15, G-coeff.=.65056

Another example offers evidence that contradicts the conventional wisdom of the Spearman-Brown formula.

H'=4, I'=5, Total n=20, G-coeff.=.71284

H'=7, I'=5, Total n=35, G-coeff.=.81288

H'=4, I'=10, Total n=40, G-coeff.=.80522

In this example, the second case has a smaller total number of items but a higher generalizability coefficient than the third case. That is, constructing the test with 7 passages containing 5 items is more efficient than using 4 passages containing 10 items per passage. It is not unusual in G-theory, when more than one facet is involved, that an increased number of items does not guarantee a higher reliability

estimate. That is, the relationship in the Spearman-Brown formula does not hold up in this situation.

The relationship of the passage effect with the item-within-passage effect is more evident in Table 5 and Figure 2. Here there are two situations with the same total number of items but with varying numbers of passages and different numbers of items within each passage.

Insert Table 5 About Here

Insert Figure 2 About Here

For these two situations, the generalizability coefficients of the p x I random effects D-study design, which produce exactly the same value as Cronbach's coefficient ALPHA based on items scores, were calculated. For each fixed total number of items, the ALPHA coefficients have higher values than the generalizability coefficients in each of the D-studies. This finding is consistent with the results from Table 1. However, the differential effects of passages and items within each passage can be seen from these data. That is, the difference between Cronbach's coefficient ALPHA (p x I design) and the generalizability coefficient goes down as the number of passages goes up (with a fixed number of items within each passage). There is an adverse effect of items within each passage. That is, the difference between Cronbach's coefficient ALPHA (p x I design) and the generalizability coefficient increases as the number of items within each passage increases. From this result, it can be inferred that increasing the number of passages is a more efficient way to obtain the desired reliability than increasing the number of items within each passage.

The underlying cause of this result can be explained in terms of within-passage and between-passage inter-item correlations also. Earlier it was shown that if the average within-passage inter-item correlation is greater than the average between-

passage inter-item correlation, a positively biased reliability estimate would result from using Cronbach's coefficient ALPHA based on item scores. At this point, it can be shown that the magnitude of this positive bias could be influenced by the number of within-passage and between-passage inter-item correlations as well as the average difference between within-passage and between-passage inter-item correlations. The average of total inter-item correlations was defined as the weighted average of the average of within-passage inter-item correlations and between-passage inter-item correlations ($\bar{\rho}_t = \frac{n_w}{n_+}\bar{\rho}_w + \frac{n_b}{n_+}\bar{\rho}_b$,where n_+ is the total number of inter-item correlations, n_w is the number of inter-item correlations within passages, and n_b is the number of inter-item correlations between passages). Therefore, the average of all inter-item correlations is influenced by an imbalance in the number of within-passage and between-passage inter-item correlations.

(1) H'=8, I'=5, Total n=40, Total r=1560, Within r=160, Between r=1400

(2) H'=5, I'=8, Total n=40, Total r=1560, Within r=280, Between r=1280

The reliability estimate of scores from the first test composed of 8 passages and 5 items per passage is less influenced by the within-passage inter-item correlations than that of the second test composed of 5 passages and 8 items per passage. The ratio of the numbers of between-passage inter-item correlations to within-passage inter-item correlations is about 8.75 for the first case and about 4.57 for the second case. That is, the second case is relatively more dominated by the within-passage inter-item correlations. If the average within-passage inter-item correlation is greater than the average between-passage inter-item correlation, a higher positive bias would be expected from using Cronbach's coefficient ALPHA in the case of the test composed of 5 passages and 8 items within each passage.

So far, the results of the p x (I : H) random effects D-study design with varying total number of items have been presented. Now results from the D-studies with a fixed total number of items are presented. This situation can be thought of as a more

realistic one because test construction is usually restricted to a fixed total number of items for a test, as determined by practical considerations. At first, the total number of items was fixed to be the same number as in the original test, and also the number of passages was fixed to be the same number of passages as the original test with varying numbers of items within each passage.

Insert Table 6 About Here

For this analysis, a reasonable range of number of items per passage was decided upon and various kinds of D-studies were completed. Only five representative combinations are presented. For a given item combination structure, the order in which items are presented within each passage is not important. That is, a given combination of items produces the same generalizability coefficient regardless of item order. For each test, the first row represents a somewhat unrealistic combination to estimate the lower bound of the reliability estimates; the last row represents the item combination having about an equal number of items per passage to produce the upper bound of the reliability estimates. The results of Table 6 provide a reasonable range of reliability estimates under these restrictions. That is, if we fixed the total number of items and the number of passages, we can expect that the variability of reliability estimates with varying numbers of items within each passage would be very small (under .01). Therefore, there is little need to be concerned about the item effect within each passage if the total number of items and the number of passages are fixed. On the basis of this result, another type of D-study, with fixed total number of items and with varying number of passages and varying number of items within each passage was completed.

Insert Table 7 About Here

The number of items within each passage was nearly equal for this analysis.

Such reasonably small variation in the number of items within each passage would produce similar reliability estimates under the restrictions of fixed total number of items and fixed number of passages. The graphical representation of Table 7 is in Figure 3.

Insert Figure 3 About Here

The passage effect and item within passage effect cannot be disentangled because the items are nested within passages and the number of passages and the number of items within each passage change simultaneously. However, it was shown earlier that the passage effect influences the reliability estimates in more dramatic ways than item-within-passage effect. Therefore, the graph in Figure 3 uses "number of passages" on the horizontal scale. There are some trends in Figure 3:

1. Grade 4 and grade 8 Reading Comprehension tests produce very similar plots.
2. Grade 4 and grade 8 Maps and Diagrams tests also produce similar plots.
3. The graph of grade 11 Test L is very similar to that for the grade 8 Maps and Diagrams test.
4. Grade 4 and grade 8 Maps and Diagrams and grade 11 Test L curves begin to flatten out sooner than the curves for grade 4 and grade 8 Reading Comprehension tests.

There are several possible explanations for these trends. First, the test content might be the underlying factor for these trends. That is, these trends might be interpreted as the interaction effect based on test content. The Reading Comprehension test and Maps and Diagrams test contain testlets that vary between tests and, differentially within tests. Therefore, tests of similar content might produce similar trends. This reasoning can help explain the first and the second trend, but this contention does not explain the third trend, the similarity between grade 11 Test L and grade 8 Maps and Diagrams tests. It is reasonable to expect that the grade 11 Test L is more similar to the Reading Comprehension tests than to the Maps

and Diagrams tests.

Second, the total number of items might be considered the underlying factor. That is, if the smaller total number of items were used, the increase of the number of passages would not much influence the number of items within each passage. The test with the smaller number of total items would reach the asymptotic point earlier than the test with more total items. So tests with similar total numbers of items would be expected to produce similar trends. This contention also can explain the first and second trends and some part of fourth trend, but not the third or some parts of the fourth.

Third, the magnitude of the variance component, $\sigma^2(pH)$, might be the underlying factor for these trends:

Grade 4 RC	estimate of $\sigma^2(pH)$ = .01599
Grade 4 M & D	estimate of $\sigma^2(pH)$ = .00752
Grade 8 RC	estimate of $\sigma^2(pH)$ = .01630
Grade 8 M & D	estimate of $\sigma^2(pH)$ = .00979
Grade 11 Test L	estimate of $\sigma^2(pH)$ = .00986

All four trends listed above can be explained with these estimated values of this variance component. That is, similar estimated variance components would be thought to produce similar trends. For example, grade 11 Test L and grade 8 Maps and Diagrams tests have similar estimated variance components (.00986 and .00979), and they produce very similar trends. In the introduction section, the important role of the estimated variance component $\sigma^2(pH)$ in p x (I : H) random effect D-study design was described. The same idea could be applied to this situation.

It appears that the relative magnitude of the variance component is a more reasonable explanation than the other two, however the influence of the two other factors should not be disregarded. Those factors could be thought of as indirect influences on the reliability trends, mediated by the relative magnitude of the

variance component. That is, those factors could be investigated as the possible variables to influence the magnitude of variance components. However, such an investigation is beyond main purposes of this study.

Finally, in a practical test construction situation, a graph like Figure 3 can be used to determine the most efficient measurement procedure. For example, if the test developer fixed the total number of items (usually a reasonable restriction), the number of passages needed to obtain the desired reliability could be determined. In the grade 4 Reading Comprehension test, for example, when .82 is the desired reliability, about six passages are needed given 44 total items. According to Table 6, if the total number of items and number of passage are fixed, the variability of the reliability estimates would be very small. Therefore, there should be little concern about the number of items per passage in the reasonable range of items within each passage. For another example, for grade 8 Maps and Diagrams test, if .85 is the desired reliability, it would be necessary to increase the total number of items because it would not be possible to get the desired reliability with only 33 total items. In this case, another D-study would be needed to determine the most efficient measurement procedure to obtain the desired reliability.

Conclusions

This study provides another way of gathering information with a G-theory approach for evaluating the reliability of testlet-based test scores. The p x (I : H) completely random effects design with unequal numbers of items within passages was used. The conclusions based on the results of this study are:

First, the present study provides empirical evidence that Cronbach's coefficient ALPHA based on item scores leads to positively biased reliability estimates of test scores composed of testlets.

Second, the empirically estimated magnitude of overestimation is about .04.

Third, the within-passage inter-item correlations and the between-passage inter-item correlations have different distributional characteristics, especially in their location statistics. This study provides empirical evidence that the number of within-passage and between-passage inter-item correlations as well as the difference between average within-passage and between-passage inter-item correlations influence the magnitude of overestimation using Cronbach's coefficient ALPHA based on item scores. The use of passage scores in this situation is reasonable.

Fourth, manipulating the number of passages is a more productive way to obtain efficient measurement procedures than is manipulating the number of items within each passage. Given a fixed total number of items and a fixed number of passages, the variability of reliability estimates with varying numbers of items per passage is small (under 0.01). Test constructors and publishers should not have much concern about the distribution of items to passages under these restrictions.

Fifth, some trends can be found in the reliability estimates ploted against the number of passages. The magnitude of the estimated variance component $\sigma^2(pH)$, representing person by passage interaction in a D-study, can explain these trends. However, the mediated effects of the types of tests (or content) and the total number of items should not be disregarded.

References

American Psychological Association.(1985). *Standards for educational and psychological testing.* Washington, DC : Author.

Anastasi, A.(1988). *Psychological testing(6th ed.).* New York : Macmillan.

Brennan, R.L.(1992). *Elements of generalizability theory.* Iowa City, Iowa : ACT.

Brennan, R.L.(1994). Variance components in generalizability theory. In C.R. Reynolds(Ed.). *Cognitive assessment : A multidisciplinary perspective.* New York: Pleum Press.

Brennan, R.L., Jarjoura, D., & Deaton, E.L.(1980). *Some issues concerning the estimation and interpretation of variance components in generalizability theory.* (ACT Technical Bulletin, No.36). Iowa City, Iowa : ACT.

Cronbach, L.J., Gleser, G.C., Nanda, H., & Rajaratnam, N.(1972). *The dependability of behavioral measurements : Theory of generalizability for scores and profiles.* New York : Wiley.

Cronbach, L.J., Linn, R.L., Brennan, R.L., & Haertel, E.(1995). *Generalizability analysis for educational assessments(Evaluation Comments).* Los Angeles : University of California, Center for Research on Evaluation, Standards, and Student Testing.

Feldt, L.S. & Brennan, R.L.(1989). Reliability. In R.L. Linn(Ed.) *Educational measurement.* Washington, DC : American Council on Education, 105-146.

Feldt, L.S., Forsyth, R.A., Ansley, T.N. & Alnot, S.D.(1994). *Iowa Tests of Educational Development : Interpretive guide for teachers and counselors.* Chicago, Illinois : The Riverside Publishing Company.

Frisbie, D.A. & Druva, C.A.(1986). Estimating the reliability of multiple true-false tests. *Journal of Educational Measurement.* 23(2), 99-105.

Hoover, H.D., Hieronymus, A.N., Frisbie, D.A., & Dunbar, S.B.(1994) *Iowa Tests of Basic Skills : Interpretive guide for school administrators.* Chicago, Illinois : The Riverside Publishing Company.

Jarjoura, D & Brennan, R.L.(1981). *Three variance components models for some measurement procedures in which unequal numbers of items fall into discrete categories.* (ACT Technical Bulletin, No. 37).

Lawrence, I.M.(1995). *Estimating reliability for tests composed of item sets.* (ETS Research Report) Princeton, New Jersey : ETS.

Sireci, S.G., Thissen, D. & Wainer, H.(1991). On the reliability of testlet-based tests. *Journal of Educational Measurement.* 28(3), 237-247.

Thorndike, R.L.(1953). Reliability. In E.F. Lindquist(Ed.) *Educational measurement.* Washington, DC : American Council on Education. 560-620.

Wainer, H. & Kiely, G.L.(1987). Item clusters and computerized adaptive testing : A case for testlets. *Journal of Educational Measurement,* 24(3), 185-201.

Wainer, H. & Lewis, C.(1990). Toward a psychometrics for testlets. *Journal of Educational Measurement,* 27(1), 1-14.

Wainer, H. & Thissen, D.(1996). How is reliability related to the quality of test scores? What is the effect of local dependence on reliability?, *Educational Measurement : Issues and Practice,* 15(1), 22-29.

Table 1

Descriptive Statistics for Data Sources Used in This Study

Test	Sample Size	n	n(h)	n(i:h)	mean	S.D.	sk	ku
ITBS Gr.4 RC	3,032	44	9	6,5,3,2,6,5,5,6,6	24.2	8.64	-0.065	2.093
ITBS Gr.4 M & D	3,003	26	4	6,6,7,7	16.5	5.43	-0.357	2.260
ITBS Gr.8 RC	3,074	57	9	8,3,6,6,5,6,8,3,12	29.1	12.28	0.317	2.082
ITBS Gr.8 M & D	3,007	33	5	7,7,6,6,7	16.8	6.44	0.187	2.198
ITED Gr.11 Test L	2,919	44	5	9,8,8,9,10	27.3	10.24	-0.271	1.885

Notes :

RC : Reading Comprehension
M & D : Maps and Diagrams
Test L : Literary Materials
Sample Size : number of examinees
n : number of items in a test
n(h) : number of passages in a test
n(i:h) : number of items within each passage
mean : sample mean of the raw scores
S.D. : standard deviation of the raw scores
sk : skewness of the raw score distribution
ku : kurtosis of the raw score distribution

Table 2

Reliability Estimates of Item-Score Coefficient ALPHA and Passage-Score Coefficient ALPHA, and Generalizability Coefficient of the p x (I : H) Random Effects Design

Test	Item α (A)	G - Coeff. (B)	Passage α (C)	Difference (A-B)	Difference (A-C)	Difference (B-C)
ITBS Gr.4 RC	.890	.848	.837	.042	.053	.011
ITBS Gr.4 M & D	.844	.805	.800	.039	.044	.005
ITBS Gr.8 RC	.928	.888	.869	.040	.059	.019
ITBS Gr.8 M & D	.839	.794	.793	.045	.046	.001
ITED Gr.11 Test L	.926	.892	.887	.034	.039	.005
Average				.0400	.0482	.0082

Table 3

Distribution Statistics for Within-Passage Inter-item Correlations and Between-Passage Inter-item Correlations

Test	n	mean	mean diff.	mean ratio	S.D.	sk	ku	range
ITBS Gr.4 RC								
Within	94	.225	.077	1.52	.087	.172	2.207	.064 - .404
Between	852	.148			.065	.280	2.694	-.007 - .353
ITBS Gr.4M & D								
Within	72	.204	.038	1.23	.071	.375	3.794	.054 - .442
Between	253	.166			.049	.132	2.596	.047 - .291
ITBS Gr.8 RC								
Within	183	.248	.072	1.41	.085	.262	2.875	.065 - .525
Between	1413	.176			.056	.484	3.419	.037 - .397
ITBS Gr.8M & D								
Within	93	.170	.042	1.33	.071	.068	2.508	.027 - .331
Between	435	.128			.050	-.087	2.536	.006 - .264
ITED Gr.11TestL								
Within	173	.266	.050	1.23	.081	.211	2.753	.078 - .258
Between	773	.216			.061	.369	3.202	.035 - .433
Average			.056	1.34				
Within		.223						.058 - .392
Between		.167						.024 - .348

Notes :

n	: number of inter-item correlations
mean	: sample mean of the inter-item correlations
S.D.	: standard deviation of the inter-item correlations
sk	: skewness of the inter-item correlaiton distribution
ku	: kurtosis of the inter-item correlation distribution
range	: range from the lowest inter-item correlation to the highest

Table 4

Generalizability Coefficients of the p x (I : H) Random Effects D-Study Design of the ITBS Gr.8 M & D Test with Varying Number of Passages and Number of Items Per Passage

H' \ I'	4	5	6	7	8	9	10
3	.608	.651	.682	.707	.727	.743	.756
4	.674	.713	.741	.763	.780	.794	.805
5	.721	.756	.782	.801	.816	.828	.838
6	.756	.788	.811	.828	.842	.852	.861
7	.784	.813	.834	.849	.861	.871	.879
8	.805	.832	.851	.865	.876	.885	.892

Notes :

I' : number of items within each passage in a D-Study

H' : number of passages in a D-Study

Table 5

Generalizability Coefficients of the p x (I : H) and the p x I Random Effects Designs of ITBS Gr.8 M & D Test with Varying Total Number of Items.

Total number of items	p x (I : H) Design (A)			p x I Design (B)		Difference between G-Coeff. (B-A)
	H'	*I'*	G-Coeff.	*I"*	G-Coeff.	
20	4	5	.713	20	.759	.046
	5	4	.721			.038
25	5	5	.756	25	.798	.042
	5	5	.756			.042
30	6	5	.788	30	.826	.038
	5	6	.782			.044
35	7	5	.813	35	.847	.034
	5	7	.801			.046
40	8	5	.832	40	.863	.031
	5	8	.816			.047
45	9	5	.848	45	.876	.028
	5	9	.828			.048
50	10	5	.861	50	.887	.026
	5	10	.838			.049

Notes :

I' : number of items within each passage in a p x (I : H) D-Study

H' : number of passages in a p x (I : H) D-Study

I" : number of items in a p x I D-Study

Table 6

Generalizability Coefficients of the p x (I : H) Random Effects Design with Fixed Total Number of Items and Fixed Number of Passages in a Test and Varying Number of Items within Each Passage.

Test	Total n	Fixed *H'* Varying *I'*	G-Coeff.	Reasonable Range
ITBS Gr.4 RC	44	3,3,3,3,3,7,7,7,8	.844	.844-.851
		2,3,3,4,5,6,6,7,8	.845	(.007)
		3,3,4,4,4,5,6,7,8	.847	
		2,3,5,5,5,6,6,6,6	.848	
		4,5,5,5,5,5,5,5,5	.851	
ITBS Gr.4 M & D	26	3,3,10,10	.796	.796-.805
		4,4,9,9	.801	(.009)
		4,5,7,10	.801	
		5,6,6,9	.804	
		6,6,7,7	.805	
ITBS Gr.8 RC	57	2,3,3,4,5,9,10,10,11	.884	.884-.894
		3,3,5,6,6,6,8,8,12	.888	(.010)
		3,4,5,5,6,7,8,9,10	.890	
		4,5,5,6,6,7,8,8,8	.892	
		6,6,6,6,6,6,7,7,7	.894	
ITBS Gr.8 M & D	33	4,4,5,9,11	.786	.786-.794
		3,7,7,8,8	.791	(.008)
		4,6,6,8,9	.791	
		5,6,7,7,8	.793	
		6,6,7,7,7	.794	
ITED Gr.11 Test L	44	5,5,6,14,14	.884	.884-.892
		4,6,8,12,14	.886	(.008)
		6,7,8,11,12	.890	
		8,8,9,9,10	.892	
		8,9,9,9,9	.892	

Notes :

Total n : total number of items in a test

I' : number of items within each passage in a D-Study

H' : number of passages in a D-Study

Table 7

Generalizability Coefficients of the p x (I : H) Random Effects Design with Fixed Total Number of Items and Varying Number of Passages and Varying Number of Items within Each Passage.

Test	Total n	H'	I'	G-Coeff.
ITBS Gr.4 RC	44	2	22,22	.733
		3	14,15,15	.779
		4	11,11,11,11	.805
		5	8,9,9,9,9	.821
		6	7,7,7,7,8,8	.832
		7	6,6,6,6,6,7,7	.840
		8	5,5,5,5,6,6,6,6	.846
		9	4,5,5,5,5,5,5,5,5	.851
		10	4,4,4,4,4,4,5,5,5,5	.855
ITBS Gr.4 M & D	26	2	13,13	.772
		3	8,9,9	.794
		4	6,6,7,7	.805
		5	5,5,5,5,6	.812
		6	4,4,4,4,5,5	.817
		7	3,3,4,4,4,4,4	.821
		8	3,3,3,3,3,3,4,4	.823
		9	2,3,3,3,3,3,3,3,3	.825
		10	2,2,2,2,3,3,3,3,3,3	.827
ITBS Gr.8 RC	57	2	28,29	.786
		3	19,19,19	.829
		4	14,14,14,15	.852
		5	11,11,11,12,12	.866
		6	9,9,9,10,10,10	.876
		7	8,8,8,8,8,8,9	.884
		8	7,7,7,7,7,7,7,8	.889
		9	6,6,6,6,6,6,7,7,7	.894
		10	5,5,5,6,6,6,6,6,6,6	.897

Table 7.
(Continued)

Test	Total n	*H'*	*I'*	G-Coeff.
ITBS Gr.8 M & D	33	2	16,17	.737
		3	11,11,11	.767
		4	8,8,8,9	.784
		5	6,6,7,7,7	.794
		6	5,5,5,6,6,6	.800
		7	4,4,5,5,5,5,5	.805
		8	4,4,4,4,4,4,4,5	.809
		9	3,3,3,4,4,4,4,4,4	.812
		10	3,3,3,3,3,3,3,4,4,4	.814
ITED Gr.11 Test L	44	2	22,22	.844
		3	14,15,15	.870
		4	11,11,11,11	.884
		5	8,9,9,9,9	.892
		6	7,7,7,7,8,8	.898
		7	6,6,6,6,6,7,7	.902
		8	5,5,5,5,6,6,6,6	.905
		9	4,5,5,5,5,5,5,5,5	.907
		10	4,4,4,4,4,4,5,5.5,5	.909

Notes :

Total n : total number of items in a test
I' : number of items within each passage in a D-Study
H' : number of passages in a D-Study

Figure 1

The Passage Effect and Item-Within-Passage Effect on Generalizability Coefficients

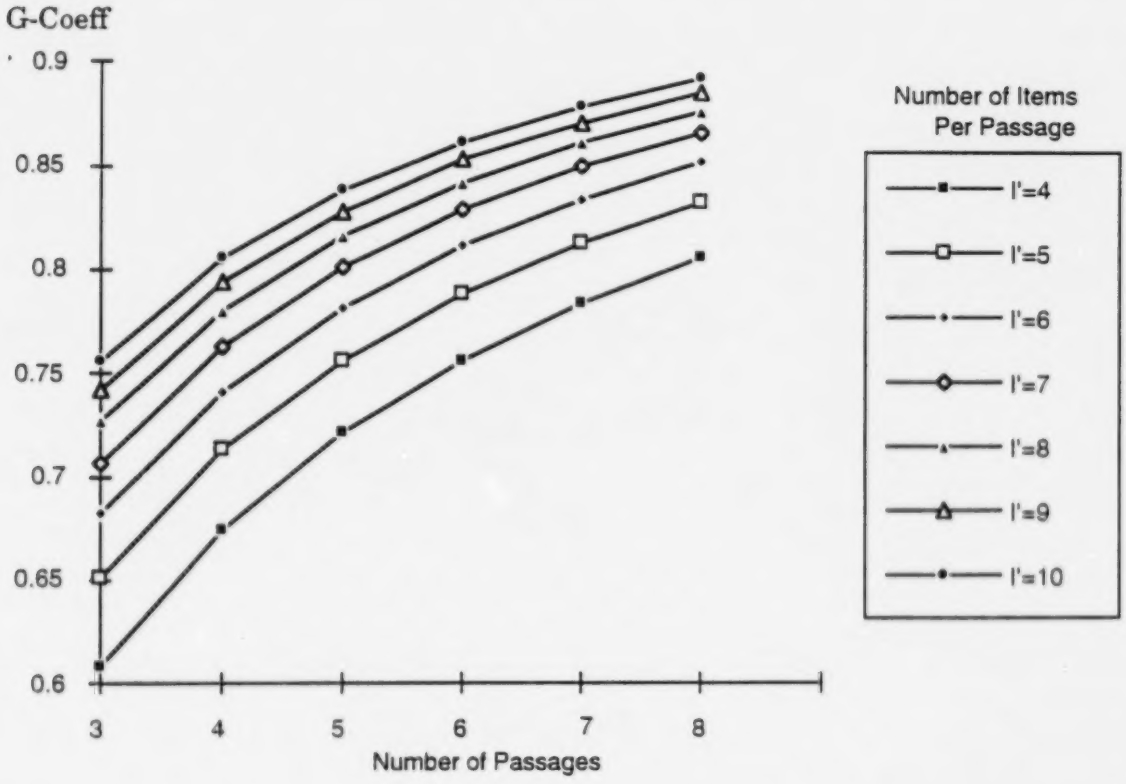


Figure 2

The Passage Effect and Item-Within-Passage Effect on Generalizability Coefficients when Total Number of Items is Fixed

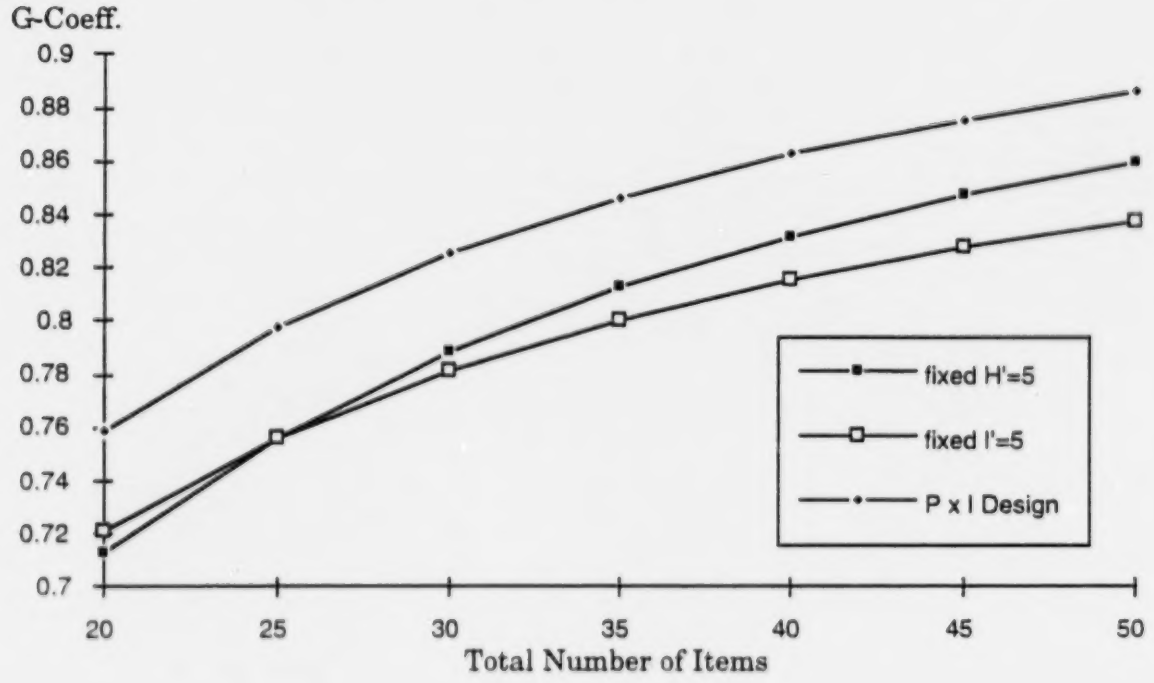
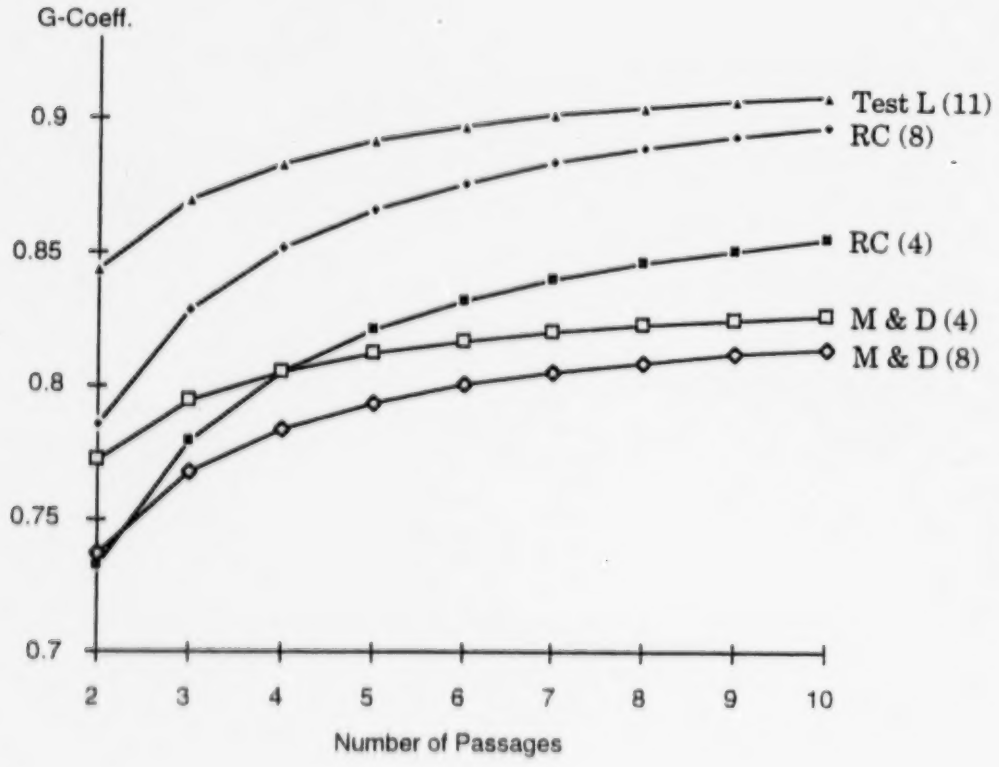


Figure 3

The Confounded Effect of Passages and Items within Passage on Generalizability Coefficients with Fixed Total Number of Items



Note

1) We appreciate the assistance of Dr. Robert Brennan in using the application program to run our data.